# Optimizing Explanations by Network Canonization and Hyperparameter Search

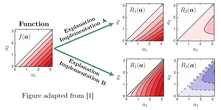
Frederik Pahde, Galip Ümit Yolcu, Alexander Binder, Wojciech Samek, Sebastian Lapuschkin

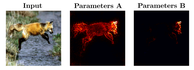
Some of the presented works have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

## Introduction

- High-quality explanations required for safety-critical ML applications
- Despite huge popularity of rule-based and modified backpropagation eXplainable AI (XAI) methods, they often **struggle with modern model architectures** with innovative building blocks
- Two main problems can be identified:
  1) **Implementation Invariance:** two models implementing the same function should lead to identical explanations ⟶ violated by many XAI methods [1]
  2) **Parameterization**: Many XAI parameter choices, optimization thereof often neglected ⟶ non-optimal explanations



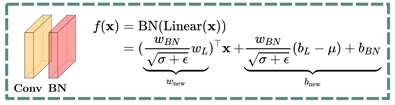

- Our contributions
  1) **Network Canonization** for modern architectures to address implementation invariance
  2) **XAI Hyperparameter Search** for optimal parameterization wrt. various XAI targets
  3) Multi-metric **XAI Evaluation Framework** to measure improvements

## Network Canonization

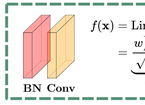
- **Problem**: Concatenation of linear operations hurts implementation invariance, e.g., Linear layer + BatchNorm (BN) at test time.
- Rule-based and modified backpropagation-based approaches are **affected by model structure**, not only the function
- **Solution**: Re-structure model such that no neighboring linear operations exist without changing the function ⟶ Network Canonization [2]
- BN parameters can be merged into neighboring Linear/Conv layers

$$\mathrm{BN}(\mathbf{x}) = w_{BN}^{\top}\left(\frac{\mathbf{x}-\mu}{\sqrt{\sigma+\epsilon}}\right) + b_{BN}$$

$w_{BN}$ := BN weights, $b_{BN}$ := BN bias, $\mu$ := running mean, $\sigma$ := running variance, $\epsilon$ := stabilizer, $\mathbf{x}$ := input sample, $w_L$ := layer weights, $b_L$ := layer bias

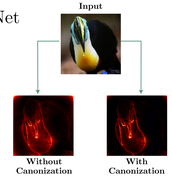



$$f(\mathbf{x}) = \mathrm{BN}(\mathrm{Linear}(\mathbf{x})) = \underbrace{\left(\frac{w_{BN}}{\sqrt{\sigma+\epsilon}}w_L\right)^{\top}}_{w_{\mathrm{new}}}\mathbf{x} + \underbrace{\frac{w_{BN}}{\sqrt{\sigma+\epsilon}}(b_L-\mu)+b_{BN}}_{b_{\mathrm{new}}}$$




$$f(\mathbf{x}) = \mathrm{Linear}(\mathrm{BN}(\mathbf{x})) = \underbrace{\frac{w_L^{\top}w_{BN}}{\sqrt{\sigma+\epsilon}}}_{w_{\mathrm{new}}}\mathbf{x} + \underbrace{\frac{w_L^{\top}w_{BN}\mu}{\sqrt{\sigma+\epsilon}}+w_L^{\top}b_{BN}+b_L}_{b_{\mathrm{new}}}$$

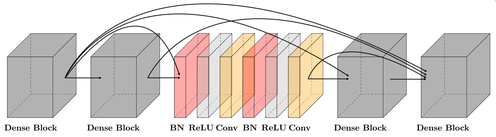
- Straightforward for simple architectures, e.g., VGG, ResNet and EfficientNet


- Complicated for highly interconnected model components, such as Dense Blocks

## DenseNet Canonization

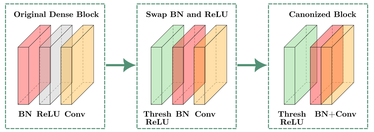
- DenseNets are **highly interconnected**: Conv layers flow into multiple BN layers and BN layers get inputs from multiple Convs
- Merging Convs/BNs is non-trivial




- **Solution:**
  1. Swap order of BN and modified ReLU
  2. Merge BN into following Conv layer

$$\mathrm{ThreshReLU}(\mathbf{x}) = \begin{cases} \mathbf{x} & \text{if } (w_{\mathrm{BN}} > 0 \text{ and } \mathbf{x} > z) \\ \mathbf{x} & \text{if } (w_{\mathrm{BN}} < 0 \text{ and } \mathbf{x} < -z) \\ z & \text{otherwise} \end{cases}$$

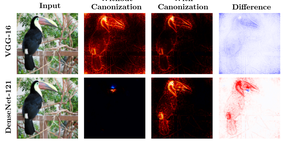
$$\text{with} \quad z = \mu - \frac{b_{\mathrm{BN}}}{w_{\mathrm{BN}}/\sqrt{\sigma+\epsilon}}$$




- Alternative: Create multiple copies of linear/conv layers (one per BN layer) ⟶ memory-consuming

## Canonization Experiments

- XAI Methods: Excitation Backprop (EB), Layer-wise Relevance Propagation (LRP)
- Dataset: ILSVRC2017 (50 random classes)
- XAI Evaluation Metrics (computed with Quantus [3]):
  1) **Complexity:** How concise are explanations?
  2) **Faithfulness**: How faithful do explanations represent the features used by the model for its prediction?
  3) **Localization:** How well do explanations localize the object of interest?
  4) **Robustness:** How robust are explanations wrt. small input perturbations?
  5) **Randomization:** How sensitive are explanations wrt. model randomization?




| Model | canonized | ↑ Complexity no | ↑ Complexity yes | ↑ Faithfulness no | ↑ Faithfulness yes | ↑ Local. (RRA) no | ↑ Local. (RRA) yes | ↑ Local. (RMA) no | ↑ Local. (RMA) yes | ↓ Robustness no | ↓ Robustness yes | ↓ Random. no | ↓ Random. yes |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| VGG-16 | EB | 0.57 | **0.59** | 0.35 | **0.36** | 0.70 | **0.71** | 0.68 | **0.70** | 0.22 | **0.18** | 1.00 | 1.00 |
| | LRP-$\alpha2\beta1$ | 0.70 | **0.84** | 0.38 | **0.39** | 0.63 | **0.67** | 0.65 | **0.77** | **0.31** | 0.34 | **0.59** | 0.66 |
| | LRP-$\varepsilon+$ | 0.51 | **0.62** | 0.36 | **0.39** | 0.69 | **0.71** | 0.64 | **0.71** | **0.19** | 0.21 | 0.57 | **0.54** |
| ResNet-18 | EB | 0.55 | **0.57** | 0.29 | 0.29 | 0.68 | **0.69** | 0.66 | **0.67** | 0.16 | **0.14** | 0.97 | 0.97 |
| | LRP-$\alpha2\beta1$ | 0.67 | **0.76** | 0.32 | 0.32 | 0.65 | **0.67** | 0.69 | **0.75** | **0.21** | 0.26 | 0.65 | **0.61** |
| | LRP-$\varepsilon+$ | 0.51 | **0.58** | 0.30 | 0.30 | 0.69 | **0.70** | 0.65 | **0.69** | **0.14** | 0.15 | 0.70 | 0.70 |
| EfficientNet-B0 | EB | **0.85** | 0.70 | 0.24 | **0.27** | **0.73** | 0.67 | **0.79** | 0.72 | 0.42 | **0.33** | **0.99** | 1.00 |
| | LRP-$\alpha2\beta1$ | 0.75 | **0.77** | **0.29** | 0.20 | **0.72** | 0.65 | **0.79** | 0.73 | **0.48** | 0.49 | 0.57 | **0.51** |
| | LRP-$\varepsilon+$ | 0.50 | **0.73** | 0.28 | **0.30** | 0.75 | 0.75 | 0.69 | **0.79** | **0.12** | 0.21 | **0.61** | 0.65 |
| DenseNet-121 | EB | **0.66** | 0.62 | 0.15 | **0.31** | 0.58 | **0.72** | 0.53 | **0.73** | 0.57 | **0.17** | **0.75** | 0.89 |
| | LRP-$\alpha2\beta1$ | **0.82** | 0.81 | 0.25 | **0.33** | 0.64 | **0.71** | 0.68 | **0.81** | 0.65 | **0.28** | **0.40** | 0.44 |
| | LRP-$\varepsilon+$ | **0.67** | 0.66 | 0.26 | **0.33** | 0.70 | **0.74** | 0.71 | **0.77** | 0.63 | **0.19** | **0.39** | 0.48 |

Results for ILVCR2017 for metrics Sparseness (Complexity), Region Perturbation (Faith-fulness), Relevance Rank Accuracy and Relevance Mass Accuracy (Localization), Avg. Sensitivity (Robustness) and Random Logit Test (Randomization). Arrows indicate whetherhigh (↑) or low (↓) are better. Best results are shown in bold.

Results for additional datasets (Pascal VOC, MS COCO) and task (VQA on CLEVR XAI) in paper

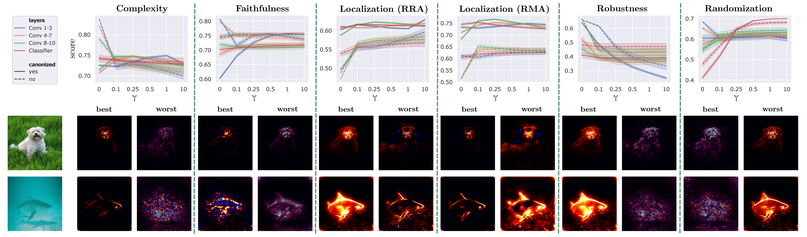
## XAI Hyperparameter Search

- **Problem**: Flexibility of XAI methods, e.g., LRP, leads to **numerous possible parameterizations** (e.g., LRP-rule/layer-assignment, rule parameters, canonization)
- **Multi-metric Hyperparameter Search**: Measure impact of different parameterizations on different metrics, estimating the explanation quality from various viewpoints
- **Experiment:**
  - VGG-13 model (with BN), pretrained on ImageNet
  - LRP $\gamma$-rule as XAI method
  - Parameters:
    1. $\gamma$ per layer (for simplicity, we group layers into Conv 1-3, 4-7, 8-10, Classifier) with $\gamma \in \{0, 0.1, 0.25, 0.5, 1, 10\}$
    2. canonization yes/no

$$R_j = \sum_k \frac{a_j \cdot (w_{jk} + \gamma\, w_{jk}^{+})}{\sum_j a_j \cdot (w_{jk} + \gamma\, w_{jk}^{+})} \cdot R_k$$

$\gamma$=0: negative/positive contributions are treated equally

$\gamma \to \infty$: negative contributions are neglected




## Conclusions

- Introduced **model canonization for DenseNet and EfficientNet**
- Impact of canonization depends on the model architecture and XAI evaluation metric, but **canonization provides an extra option to tune explanations** for the task at hand
- XAI hyperparameter choices are key. They can be tuned using our multi-metric evaluation framework. **Optimal hyperparameter choice depends on the given task**.
- Future work:
  - extend canonization to other relevant model architectures, including Vision Transformers
  - optimze hyperparameter search beyond grid search with multi-target objective

## References

[1] Montavon, Grégoire. "Gradient-based vs. propagation-based explanations: An axiomatic comparison." Explainable ai: Interpreting, explaining and visualizing deep learning (2019): 253-265.

[2] Motzkus, Franz, et al. "Measurably stronger explanation reliability via model canonization." IEEE International Conference on Image Processing (2022).

[3] Hedström, Anna, et al. "Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond." Journal of Machine Learning Research 24.34 (2023): 1-11.